Application of regression analysis in the experiment of ethanol coupling preparation of C4 olefin

Shucheng Zhong, Xuanning Wang, Qiang Chen*

School of Mathematics and Physics, Yancheng Institute of Technology, Yancheng, Jiangsu, 224003, China
**Corresponding author: chq@ycit.cn*

Abstract: *In complex scientific experiments, regression analysis has important value in data processing and analysis. Taking the preparation of C4 olefin by ethanol coupling as an example, this project established a regression model based on correlation analysis to solve the problem of the relationship between ethanol conversion, C4 olefin selectivity and temperature and time, and obtained that the goodness of fit of the quadratic fitting of ethanol conversion, C4 olefin selectivity and temperature was higher than that of the primary fitting, with an average of 0.03 higher. The selectivity of C4 olefin fluctuates between 36% and 41%, indicating that the selectivity of C4 olefin has little correlation with time under a given catalyst combination at 350℃. The ethanol conversion rate decreases with the increase of time, and finally tends to a constant value. According to the requirements of "Ethanol coupling preparation of C4 olefin" in Question B of the 2021 National Mathematical Competition in Modeling for College Students and the data provided in Annex 1 and Annex 2 of the question, this project establishes a multiple linear regression model and a single objective optimization model based on nonlinear regression to conduct a comprehensive study on ethanol conversion and maximum yield of C4 olefin.*

Keywords: *correlation analysis; Multiple linear regression; Single objective optimization model; Genetic algorithm*

1. Introduction

C4 olefin is widely used in the chemical industry and pharmaceutical production, and in the past, fossil fuels or syngas were used as feedstocks to prepare C4 olefin in traditional manufacturing processes. The use of clean energy ethanol to prepare C4 olefins has become a better choice.

How to reduce the waste of raw materials and improve the conversion rate of ethanol has become an important issue in the preparation of C4 olefin. Through experimental research, it is found that the conversion rate of ethanol and the selectivity of C4 olefin can be improved by changing the reaction temperature and adjusting the components of ethanol catalyst. The research object of this analysis comes from the question B of the 2021 National Mathematical Contest in Modeling for College students.

In the performance data table given by the title, there are 21 catalyst combinations. For each catalyst combination, the conversion rate of reactants and the selectivity data of various products at 5-7 temperatures are respectively given. Considering the large and complex experimental data, there are many factors affecting the conversion of reactants and the selectivity of products. Therefore, we choose to use the method of regression analysis to analyze and study.

Based on the data characteristics, this paper mainly studies the following issues: Analyze the influence of different catalyst combinations and temperature on ethanol conversion and C4 olefin selectivity;

To simplify the problem, the following assumptions are proposed:

1) The catalyst is assumed to be stable during the reaction;

2) Assume that the reaction is not reversible; ·

3) It is assumed that the experimental equipment is intact and there will be no air leakage or damage during the experiment.

2. Temperature effect and time effect

2.1 Data preprocessing

(1) Removal of outliers

We found that quartz sand rather than HAP was used in the A11 catalyst combination. After confirming that it was an interference sample, we decided to remove 5 groups of samples of A11.

(2) Processing missing values

We observed and analyzed the data in Annex 1 and found that the test temperature under each catalyst combination was not exactly the same. For example, the catalyst combination A6 and B1 lacked data at 325℃, while the catalyst combination A3 had data at 450℃, which would bring inconvenience to the subsequent analysis. Therefore, we adopted cubic spline interpolation. The temperature was set between 250 ° C and 400 ° C and the interval was 25 ° C, where the missing data points were interpolated and completed.

Cubic spline interpolation [1-2] formula:

$$S_i\left(x\right)=a_i+b_i\left(x-x_i\right)+c_i\left(x-x_i\right)^2+d_i\left(x-x_i\right)^3, i=0,1,...,n-1 \tag{1}$$

There are n intervals, that is, n segment functions $S_0 \sim S_{n-1}$, with n+1 nodes. The completed data obtained after interpolation calculation is shown in the following table:

Table 1: Temperature interpolation data of each catalyst combination

	250℃	275℃	300℃	325℃	350℃	375℃	400℃
A1	2.07	5.85	14.97	19.68	36.8	89.79	/
A2	4.6	17.2	38.92	56.38	67.88	74.64	/
…	…	…	…	…	…	…	…
A14	2.5	5.3	10.2	16.03	24	35.92	53.6
B1	1.4	3.4	6.7	11.8	19.3	29.7	43.6
…	…	…	…	…	…	…	…
B7	4.4	7.9	11.7	17.8	30.2	48.1	69.4

In Table 1, there are 7 different temperature levels under each catalyst combination, increasing the number of data points, which can effectively improve the accuracy of model fitting in question 1.

(3) Data transformation

The root of the variable ethanol conversion Y_1, C4 olefin selectivity Y_2, and C4 olefin yield $\sqrt{Y_1Y_2}$ is transformed by Logit, and the new variable is:

$$\begin{cases} y_1=\log(Y_1/(100-Y_1)) \\ y_2=\log(Y_2/(100-Y_2)) \\ y_3=\log(\sqrt{Y_1Y_2}/(100-\sqrt{Y_1Y_2})) \end{cases} \tag{2}$$

2.2 Data visualization

In order to more intuitively and conveniently analyze the relationship between ethanol conversion and C4 olefin selectivity and temperature, consider visualizing the data under various catalyst combinations, and draw the scatter plot as follows:

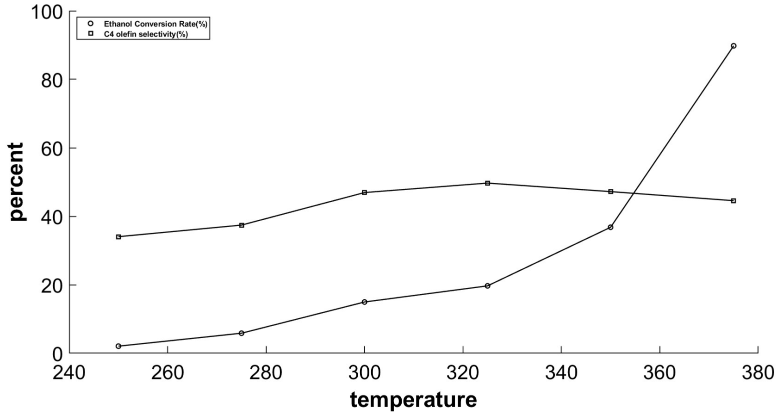

Figure 1: Scatter plot of ethanol conversion and C4 olefin selectivity with temperature in A1 team

It can be seen from Figuer 1 that the change rate of ethanol conversion with temperature increasing is fast, and the selectivity of C4 olefin with temperature increasing is relatively slow and has a downward trend. Among them, the ethanol conversion rate changes the fastest at 350℃-375℃, and the C4 olefin selectivity changes the fastest at 275℃-300℃, but the rate is much lower than the ethanol conversion rate.

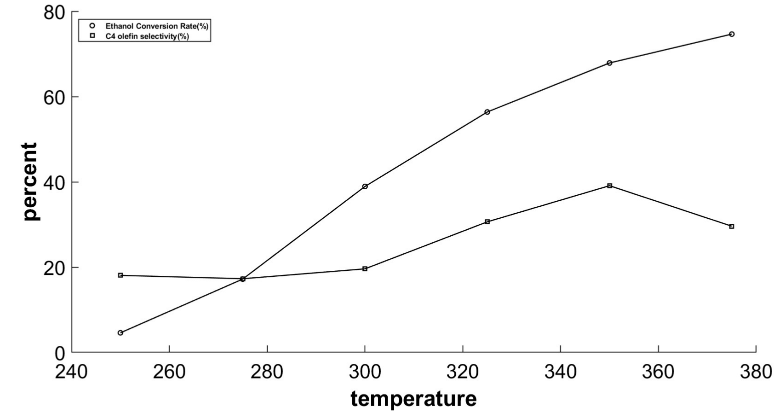

Figure 2: Scatter plot of ethanol conversion and C4 olefin selectivity with temperature in A2 team

It can be seen from Figure 2 that the change rate of ethanol conversion with temperature increasing is fast, while the selectivity of C4 olefin with temperature increasing is relatively slow and has a downward trend. Among them, the ethanol conversion rate changes the fastest at 280℃-360℃, and the C4 olefin selectivity changes the fastest at 300℃-350℃.

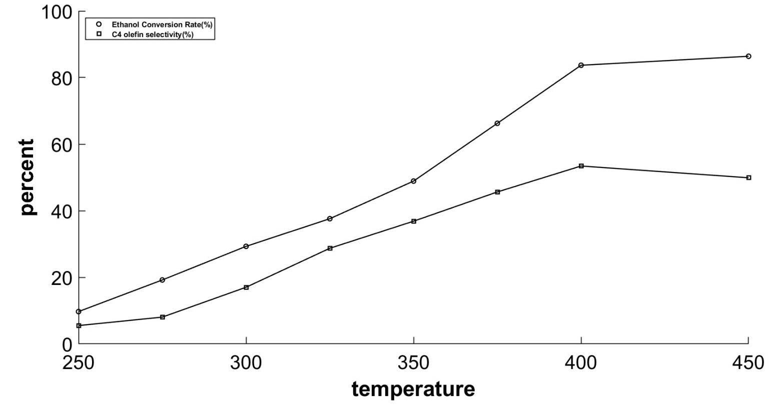

Figure 3: Scatter plot of ethanol conversion and C4 olefin selectivity with temperature in A3 team

It can be seen from Figure 3 that the selectivity of both ethanol conversion and C4 olefin increases with the increase of temperature, and the change of ethanol conversion is higher than that of C4 olefin selectivity. Among them, the ethanol conversion rate changes the fastest at 350°C-400°C, and the C4 olefin selectivity changes the fastest at 300°C-400°C.

2.3 Pearson correlation coefficients of ethanol conversion

C4 olefin selectivity and temperature were calculated based on correlation analysis of Pearson correlation coefficient [3] as follows:

Table 2: Pearson correlation coefficients

ID	Ethanol conversion	Selectivity of C4 olefins	ID	Ethanol conversion	Ethanol conversion
A1	0.9577	0.7365	A12	0.9966	0.9991
A2	0.9663	0.8386	A13	0.9995	0.9697
A3	0.9846	0.9317	A14	0.9975	0.9958
A4	0.9942	0.9698	B1	0.9966	0.9979
A5	0.9562	0.9736	B2	0.9921	0.9952
A6	0.9886	0.9388	B3	0.9964	0.9777
A7	0.9978	0.9935	B4	0.9961	0.9057
A8	0.9984	0.9965	B5	0.9965	0.9962
A9	0.9922	0.9829	B6	0.9919	0.9782
A10	0.9969	0.9002	B7	0.9926	0.9881

As can be seen from Table 2, for different catalyst combinations, the correlation between ethanol conversion and C4 olefin selectivity and temperature is large, and the Pearson correlation coefficient of 18 out of 20 data groups is greater than 0.9. The correlation between ethanol conversion and temperature is generally higher than that between C4 olefin selectivity and temperature.

2.4 The fitting equations of ethanol conversion, C4 olefin selectivity and temperature

The data in the 20 catalyst combinations are fitted by first and second regression [4], and the fitting results are obtained as follows:

For catalyst combination A1, the fitting equation is $y = 0.0425x - 14.64$: the goodness of fit is 0.917;

The quadratic fitting equation is $y = 0.000213x^2 - 0.091x + 5.841$, the goodness of fit is 0.948;

For catalyst combination A2, the first fitting equation is $y = 0.3226x - 10.58$, and the goodness of fit is 0.934.

The quadratic fitting equation is $y = -0.000233x^2 + 0.178x - 32.9$, the goodness of fit is 0.999;

……

For catalyst combination B6, the first fitting equation is $y = 0.025x - 9.74$, the goodness of fit is 0.9839;

The quadratic fitting equation is $y = 8.1\times10^{-6}x^2 + 0.0201x - 8.7$, the goodness of fit is 0.9841;

For catalyst combination B7, the first fitting equation is $y = 0.0252x - 9.5$ and the goodness of fit is 0.985.

The quadratic fitting equation is $y = 6.47\times10^{-5}x^2 - 0.0169x - 2.83$, the goodness of fit is 0.998;

It can be seen that the goodness of fit of quadratic fitting is higher than that of primary fitting, so a quadratic regression model is established below.

2.5 Analysis of test results and time

(1) Ethanol conversion

Based on the quadratic regression model, the following quadratic regression function is obtained:

$$y = 0.0002T^2 - 0.101T + 45.1 \tag{3}$$

Among them, the goodness of fit is 0.986, which is close to 1, so it is considered that the model can better reflect the change relationship between ethanol conversion and time. In summary, the segmented function is obtained:

$$y = \begin{cases} 0.0002T^2 - 0.101T + 45.1, t < 240 \\ 29.9, t \geq 240 \end{cases} \tag{4}$$

(2) C4 olefin selectivity and other products

Since there is no obvious linear or nonlinear relationship between C4 olefin selectivity and time, considering C4 olefin selectivity and other products, the analysis results of selectivity are as follows:

Table 3: The statistical analysis of reaction products

Product type	Mean value	variance	range
C4 olefin	39.00	1.37	3.60
ethene	4.52	0.04	0.53
acetaldehyde	7.19	2.13	3.62
Aliphatic alcohol	33.64	11.94	8.84
Methyl benzaldehyde	4.14	0.55	2.22

According to the relevant statistics in Table 3, the selectivity of C4 olefin is a random variable, which is not affected by time under this condition. Among all the products, C4 olefin has the highest selectivity, but the variance is relatively small, indicating that the degree of dispersion is small in the range of 36%~41%. The second is fatty alcohol, but the degree of dispersion is large; The content of other products is very small.

3. The influence of different catalyst combinations and temperatures on ethanol conversion and C4 olefin selectivity

The following multiple linear regression equation is established model [5]:

Table 4: Summary of multiple linear regression models

model	R^2	Adjust R^2	F	p
1	0.865	0.891	171	4.33×10^{-56}
2	0.808	0.851	112	5.8×10^{-46}

According to the size of the R-square value in Table 4, the fit degree of the multiple linear regression equation can be judged, R2 = 0.871, it can be seen that the equation has a high fit.

The following two sets of regression equations are obtained:

Ethanol conversion rate:

$$y_1 = 1.867\times10^{-6} + 0.642x_1 - 0.226x_2 - 0.499x_3 + 0.326x_4 + 0.769x_5 \tag{5}$$

$x_1, x_2, x_3, x_4, x_5, y$ represent respectively the ethanol concentration, Co load, Co/SiO_2 mass, HAP mass, temperature, and ethanol conversion rate.

By comparing the absolute coefficients, the influencing factors on ethanol conversion are as follows: temperature, ethanol concentration, Co/SiO2 mass, HAP mass, Co loading capacity.

C4 olefin selectivity:

$$y_2 = 2.567\times10^{-16} + 0.072x_1 - 0.405x_2 + 1.157x_3 - 0.749x_4 + 0.742x_5 \tag{6}$$

x_1,x_2,x_3,x_4,x_5,y, respectively, ethanol concentration, Co loading capacity, Co/SiO_2 mass, HAP mass, temperature, and C4 olefin selectivity.

Similarly, the influencing factors on C4 olefin selectivity are as follows: Co/SiO2 mass, HAP mass, temperature, Co loading capacity, ethanol concentration.

4. The optimal solution of C4 olefin yield

4.1 Nonlinear regression model based on C4 olefin yield

In various nonlinear relations, can be divided into three types. The first kind can be reduced to linear relation by variable substitution; The second type is y that the function form of the nonlinear relation with the independent variable is not clear. This kind of nonlinear regression problem can be solved by multiple linear stepwise regression. The third type of nonlinear problem is one in which the functional form of the nonlinear relationship y with the independent variable is determined (only the parameters are unknown), but cannot be transformed into a linear relationship by variable transformation. Such nonlinear regression problems must be solved by complex fitting methods.

The general nonlinear regression model [6] can be written as:

$$Y=\varphi\left(x_1,x_2,\cdots,x_m,\beta_1,\beta_2,\cdots,\beta_r\right)+\varepsilon \tag{7}$$

For a given set of observations $\left(x_i,y_i\right),i=1,2,\cdots,n$, the above formula can be rewritten as $y_i=f\left(x_i,\theta\right)+\varepsilon_i$, $i=1,2,\cdots,n$ .Where y_i is the dependent variable and non-random vector $x_i=\left(x_{i1},x_{i2},\cdots,x_{ik}\right)'$ is the independent variable;

$\theta=\left(\theta_0,\theta_1,\cdots,\theta_p\right)'$ is an unknown parameter vector; ε_i is the random error term and satisfies the independent identically distributed assumption,

$$\begin{cases} E\left(\varepsilon_i\right)=0,i=1,2,\cdots,n \\ \operatorname{cov}\left(\varepsilon_i,\varepsilon_j\right)=\begin{cases}\sigma^2,i=j\\0,i\neq j\end{cases},\left(i,j=1,2,\cdots,n\right)\end{cases} \tag{8}$$

If $f\left(x_i,\theta\right)=\theta_0+x_1\theta_1+x_2\theta_2\cdots,x_p\theta_p$, then $y_i=f\left(x_i,\theta\right)+\varepsilon_i$ is the model discussed earlier, and there must be $k=p$; For nonlinear models in general, the number of parameters does not correspond to the number of independent variables, so ttt is not required $k=p$.

The nonlinear fitting gives the following equation:

$$\begin{aligned} y=11787-27.7x_1-77.5T-172.8x_3^2+0.1T^2 \\ +3.7x_2x_4+0.1x_1T-3.1x_4T+482.4x_3x_4 \end{aligned} \tag{9}$$

4.2 Single objective optimization model

Using the obtained fitting function, we can build a single objective optimization model to obtain the highest C4 olefin yield:

Decision variables: The independent variable objective function of the C4 olefin yield x_1,x_2,x_3,x_4,x_5

fitting function: The objective is that the C4 olefin yield y3 is maximum

Constraint: The value range of the 5 independent variables does not exceed the maximum range in the original catalyst combination.

(1) Co load restriction:

$$0.5 \le x_1 \le 5 \tag{10}$$

(2)ethanol concentration constraints:

$$0.3 \le x_2 \le 2.1 \tag{11}$$

(3)quality constraints:

$$33 \le x_3 \le 200 \tag{12}$$

(4)HAP quality constraints:

$$33 \le x_4 \le 200 \tag{13}$$

(5) temperature constraints:

$$250 \le T \le 400 / 250 \le T \le 350 \tag{14}$$

In summary, the following single-objective optimization model is established:

$$\max \ \ y_3 = y_1 \bullet y_2 , \tag{15}$$

$$s.t.\begin{cases} 0.5 \le x_1 \le 5 \\ 0.3 \le x_2 \le 2.1 \\ 33 \le x_3 \le 200 \\ 33 \le x_4 \le 200 \\ 250 \le T \le 400 / 250 \le T \le 350 \\ y_1 = 1.864 \times 10^{-6} + 0.642x_1 - 0.226x_2 - 0.499x_3 + 0.326x_4 + 0.769x_5 \\ y_2 = 2.567 \times 10^{-16} + 0.072x_1 - 0.405x_2 + 1.157x_3 - 0.749x_4 + 0.742x_5 \end{cases} \tag{16}$$

4.3 Genetic Algorithm Solving

Genetic algorithm [7] is an iterative algorithm that has a set of solutions at each iteration, which are initially randomly generated. At each iteration, a new set of solutions is generated by a genetic operation that simulates evolution and inheritance, and each solution is judged by an objective function, making one iteration a generation. The algorithm steps are as follows:

(1) Initialization, that is, randomly generate a symbol string group;

(2) The symbol string is evaluated based on the moderate function;

(3) Using a set of genetic operations to generate a new population of symbol strings;

(4) Repeat steps (2) and (3) until the result converges. It acts on the population of feature strings in a way that does not depend on the problem itself, and the algorithm uses only the adaptation values associated with the checked points in the search space. The solution itself completes its search by performing the same, surprisingly simple operations of copying, hybridizing, and occasionally mutating. In practical applications, genetic algorithms can search complex, highly nonlinear and multi-dimensional Spaces quickly and efficiently. Surprisingly, it does not know any information about the problem itself, nor does it understand the adaptation value measure. The key to the success of genetic algorithm lies in the design of symbol string representation and genetic operation.

Genetic algorithm itself has 4 parameters, namely population size crossover probability, mutation probability and maximum genetic algebra. Set population size because the larger the population, the easier it is to find the optimal solution. The crossover probability, with a crossover probability of 0.7, ensures the full evolution of the population. The probability of variation is 0.001, in general, the probability of variation is less, and the probability of variation is 0.001, which is more in line with the law of nature. The maximum genetic algebra is 300 to ensure full convergence of optimization results.

The results are as follows: Co load is 0.5, ethanol concentration is 0.3, Co/SiO_2 mass is 200, HAP mass is 200, temperature is 399, C4 olefin yield is 40.67%. If the temperature is lower than 350℃, the Co load is 0.5, the ethanol concentration is 0.3, the Co/SiO_2 mass is 200, the HAP mass is 199, the temperature is 349, and the C4 olefin yield is 24.35%.

5. Conclusion

In view of the influence of different catalyst combinations and temperature on ethanol conversion and C4 olefin selectivity, in order to explore and determine the influence of factors on the observed variables, this paper adopted the analysis of variance in regression analysis, established a one-way analysis of variance model, and adopted random forest for testing. The results are consistent with the results of analysis of variance, which reflects the rigor and science of our results. Based on the above processing process, it can be seen that the regression analysis model is suitable for chemical reaction, which can improve the income and reduce the cost, and it is hoped that it can be more widely applied in the future.

References

[1] Hongjie J, Yue Z, Liguo Z, et al. Stochastic pavement reconstruction method based on harmonic superposition technique of cubic spline interpolation[J]. Journal of Vibration and Shock, 2023, 42(10):23-30+143.

[2] Haiqing C, Xu Z, Leilei Z, et al. Fitting generalized logistic distribution by least squares based on the logistic transformation of order statistics[J]. Communications in Statistics - Theory and Methods, 2023, 52(2).

[3] Yuanshang Z, Weifang L. Research on typical scenarios based on Pearson correlation coefficient fusion density peak and entropy weight method[J]. Electric Power, 2023, 56(05):193-202.

[4] Feng D, Li S, Hui, D, et al. Environmental correction of log curves based on multiple linear fitting and convolutional neural networks: A case study of Paleogene-Cretaceous strata in Xinhe-Sandaoqiao area[J]. Journal of Engineering Geophysics, 2023, 20(02):253-265.

[5] Qi Y. An empirical study of influencing factors of China's foreign exchange reserves based on multiple linear regression model[J]. Theory of Chinese commerce, 2023(13):8-11.

[6] Wei W, Meng W, Jun C, et al. Establishing a dual multivariate nonlinear regression constitutive equation for EB furnace melting TC4 titanium alloy based on DMNR model [J]. Rare Metal Materials and Engineering, 2021, 50(10):3609-3620.

[7] Qiang Z, Weipao M, Qingsong L, et al. Optimization design of vertical axis wind turbine special airfoil based on multi-objective genetic algorithm[J]. Journal of Solar energy, 2023, 44(04):9-16.